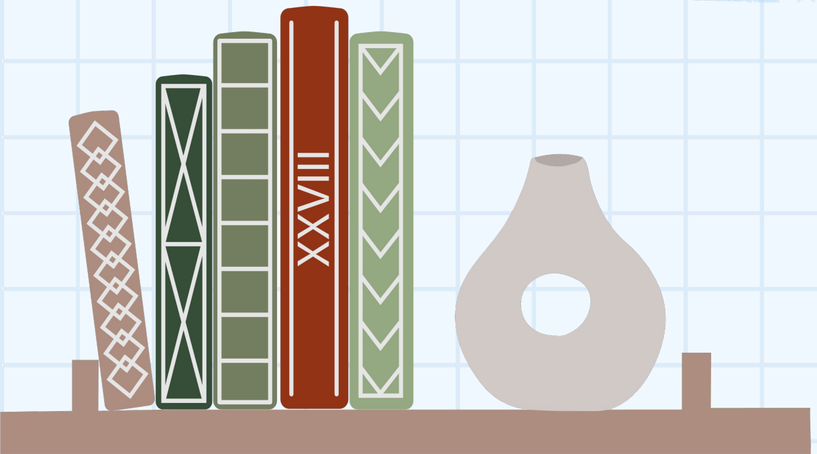
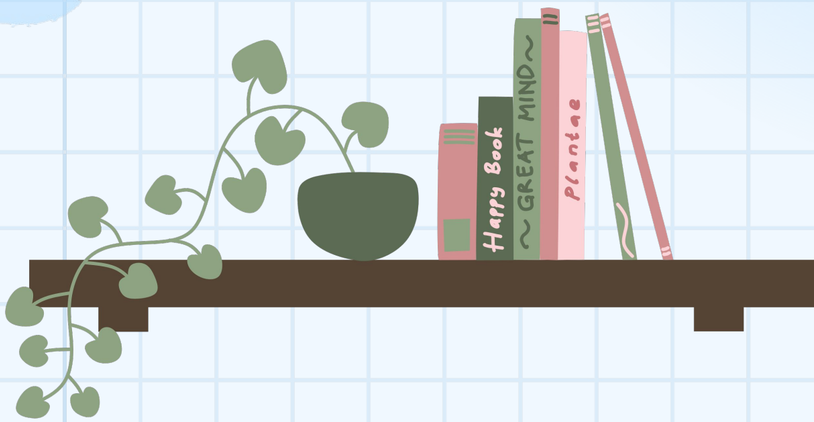
Powered by
Futurense

BS./BSC.IN

Applied AI and Data Science

Basics of Data Analytics





Let's dive into and learn:

1. Different Types of Data File Formats

Powered by Futurense

Types of Data File Formats

Powered by

Futurense

- Data can also be classified into different types based on the file formats.
- In any data analytics project, you will most likely work with different types of data formats.
- Identifying the format the data comes in is a crucial step in the exploratory phase.
- Different formats require different tools to process data.
- Different formats have their benefits and limitations.

Common Data Formats

Powered by

Futurense

1. Delimited text files e.g. csv
2. Spreadsheets e.g. Excel or Google Sheets
3. Extensible Markup Language e.g. XML
4. Portable Document Format e.g. PDF
5. Javascript Object Notation JSON

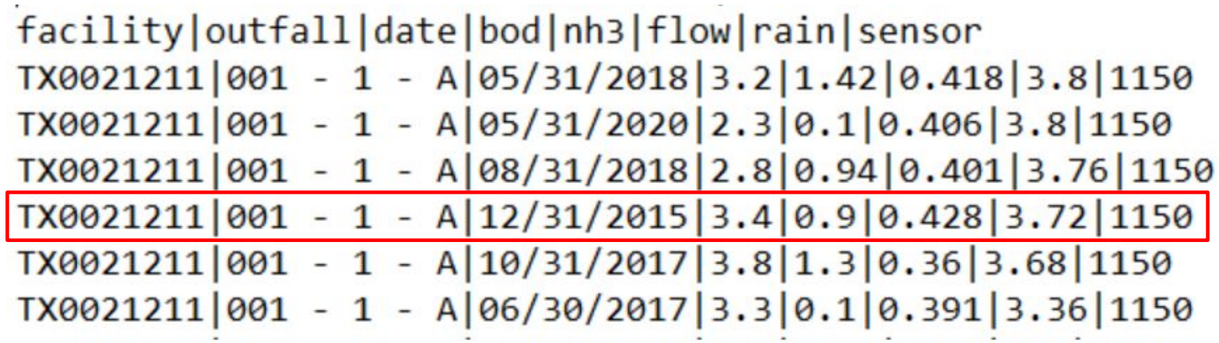
Delimited Text Files

Powered by

Futurense

- Files store data as text
- Each value is separated by a ‘**delimiter**’
- **Delimiter** – A set of one of more characters which specifies the boundaries between different values in the data
- Common ones are comma, tab, vertical bar, colon, space
- Most common: comma separated text CSV, tab separated text TSV
- Can be processed by almost all existing software and programs

Delimited Text Files



```
facility|outfall|date|bod|nh3|flow|rain|sensor
TX0021211|001 - 1 - A|05/31/2018|3.2|1.42|0.418|3.8|1150
TX0021211|001 - 1 - A|05/31/2020|2.3|0.1|0.406|3.8|1150
TX0021211|001 - 1 - A|08/31/2018|2.8|0.94|0.401|3.76|1150
TX0021211|001 - 1 - A|12/31/2015|3.4|0.9|0.428|3.72|1150
TX0021211|001 - 1 - A|10/31/2017|3.8|1.3|0.36|3.68|1150
TX0021211|001 - 1 - A|06/30/2017|3.3|0.1|0.391|3.36|1150
```

- Each row has values separated by the delimiter
- Allows values of varying length
- Standard format for tabular data

Spreadsheet Files

- In spreadsheets data are organized into neat rows and columns.
- Can be easily processed by most software and programs.
- Sometimes comes with built-in functionality that allows you to do quick statistical analysis
- E.g. Microsoft Excel or Google Sheets

Spreadsheet Files

	A	B	C	D	E	F	G	H	I	J	K
1	Order #	First Name	Last Name	Email	Country	IP address	Total	Item #	Payment	Shipping	Status
2	1	Dalton	Kramer	dalton@email.com	France	211.91.226.108	99	868	Card	Regular	In progress
3	2	Gita	Tetterton	gita@email.com	USA	222.153.179.100	99	537	Card	Regular	Delivered
4	3	Weston	Jurgens	weston@email.com	Spain	203.123.236.1	99	616	Paypal	Regular	Delivered
5	4	Brad	Chupp	brad@email.com	France	202.183.111.122	49	673	Card	Fast	Delivered
6	5	Marybeth	Baumann	marybeth@email.com	Italy	214.132.168.129	199	829	Bank	Regular	In progress
7	6	Allyson	Feder	allyson@email.com	Italy	182.108.190.85	29	40	Card	Regular	In progress
8	7	Lucile	Folks	lucile@email.com	Greece	18.64.161.62	199	548	Paypal	Fast	In progress
9	8	Mickey	Rusk	mickey@email.com	Canada	40.18.115.207	49	53	Paypal	Fast	Delivered
10	9	Clarine	Esslinger	clarine@email.com	Greece	185.134.23.86	49	817	Bank	Regular	Delivered
11	10	Kimberly	Penny	kimberly@email.com	France	34.72.165.11	99	998	Bank	Regular	In progress
12	11	Colleen	Kellough	colleen@email.com	USA	73.51.152.185	49	14	Paypal	Regular	In progress
13	12	Nettie	Edmonds	nettie@email.com	Spain	94.133.138.234	99	670	Card	Fast	Delivered
14	13	Duncan	Rickenbacker	duncan@email.com	France	211.91.226.108	199	869	Card	Regular	Delivered
15	14	Marchelle	Diedrich	marchelle@email.com	Italy	222.153.179.100	29	536	Paypal	Regular	Delivered
16	15	Mariano	Murrell	mariano@email.com	Italy	203.123.236.1	99	477	Card	Fast	Delivered

- Each row is a data entry
- Every value is already in a different cell

Extensible Markup Language

Powered by

Futurense

- It is a markup language that has specific rules for encoding data.
- A markup language is text-encoding system that defines the layout and presentation of a digital document.
- Readable by humans as well as machines.
- Platform and programming language independent.
- Makes data sharing simple across different systems.

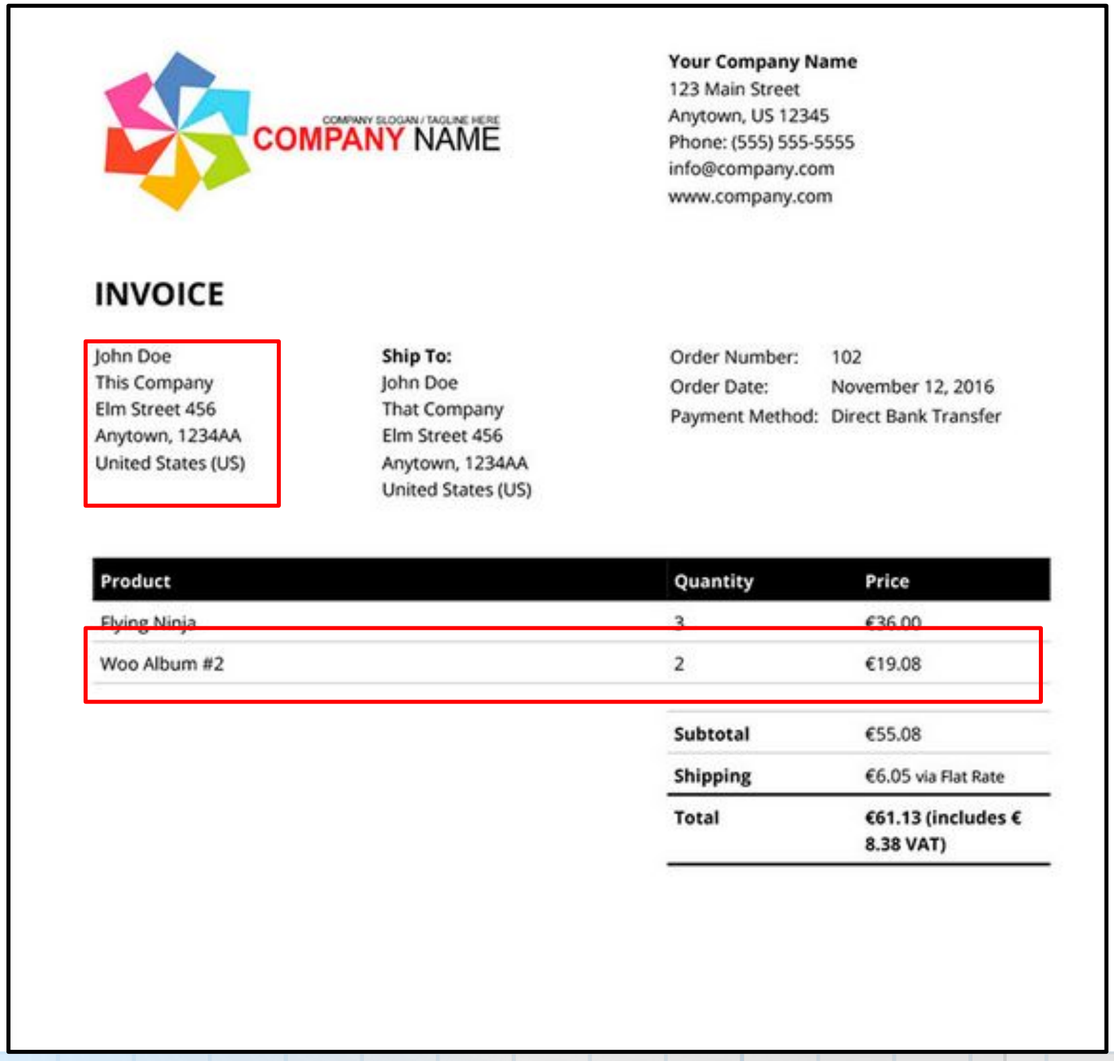
Extensible Markup Language

Portable Document Format PDF

- Portable Document Format PDF is a format developed by Adobe
- Presents document in a format independent of application and operating systems
- Displayed the same way on any device
- Often used in financial, legal documents
- Data collected through forms

Portable Document Format PDF

Powered by Futurense



COMPANY SLOGAN / TAGLINE HERE
COMPANY NAME

Your Company Name
123 Main Street
Anytown, US 12345
Phone: (555) 555-5555
info@company.com
www.company.com

INVOICE

John Doe
This Company
Elm Street 456
Anytown, 1234AA
United States (US)

Ship To:
John Doe
That Company
Elm Street 456
Anytown, 1234AA
United States (US)

Order Number: 102
Order Date: November 12, 2016
Payment Method: Direct Bank Transfer

Product	Quantity	Price
Flying Ninja	3	€36.00
Woo Album #2	2	€19.08

Subtotal	€55.08
Shipping	€6.05 via Flat Rate
Total	**€61.13 (includes € 8.38 VAT)**

Javascript Object Notation JSON

Powered by

Futurense

- Javascript Object Notation JSON is a text-based standard format
- Used for transmitting data over the web
- Language Independent data format
- Can be read by most programming languages
- Compatible across many browsers
- Can be used to share data across many types and sizes

Javascript Object Notation JSON

```
[
  {
    "Platform" : "Android",
    "Favorite Food" : "Noodle!",
    "Language" : "C#"
  },
  {
    "Platform" : "iOS",
    "Favorite Food" : "Pasta!",
    "Language" : "Swift"
  },
  {
    "Platform" : "iOS",
    "Favorite Food" : “Rice!”,
    "Language" : “Java”
  }
]
```

Powered by Futurense

Recap

Powered by Futurense

Different file formats that data can come in.

Powered by Futurense

Thank you